We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

7

Recognition of Emotion from Speech: A Review

S. Ramakrishnan
Department of Information Technology,
Dr. Mahalingam College of Engineering and Technology, Pollachi
India

1. Introduction

Emotional speech recognition is an area of great interest for human-computer interaction. The system must be able to recognize the user's emotion and perform the actions accordingly. It is essential to have a framework that includes various modules performing actions like speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions. The classifications of features involve the training of various emotional models to perform the classification appropriately. Another important aspect to be considered in emotional speech recognition is the database used for training the models. Then the features selected to be classified must be salient to identify the emotions correctly. The integration of all the above modules provides us with an application that can recognize the emotions of the user and give it as input to the system to respond appropriately.

In human interactions there are many ways in which information is exchanged (speech, body language, facial expressions, etc.). A speech message in which people express ideas or communicate has a lot of information that is interpreted implicitly. This information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people, among others. The speaker's emotional state is closely related to this information. In evolutionary theory, it is widely accepted the "basic" term to define some emotions. The most popular set of basic emotions: happiness (joy), anger, fear, boredom, sadness, disgust and neutral. Over the last years the recognition of emotions has become a multi-disciplinary research area that has received great interest. This plays an important role in the improvement of human–machine interaction. Automatic recognition of speaker emotional state aims to achieve a more natural interaction between humans and machines. Also, it could be used to make the computer act according to the actual human emotion. This is useful in various real life applications as systems for real-life emotion detection using a corpus of agent-client spoken dialogues from a medical emergency call centre, detection of the emotional manifestation of fear in abnormal situations for a security application, support of semi-automatic diagnosis of psychiatric diseases and detection of emotional attitudes from child in spontaneous dialog interactions with computer characters. On the other hand, considering the other part of a communication system, progress was made in the context of speech synthesis too. The use of bio signals (such as ECG, EEG, etc.), face and body images are an interesting alternative to detect emotional states. However, methods to record and use these signals are more invasive, complex and impossible in

certain real applications. Therefore, the use of speech signals clearly becomes a more feasible option. Good results are obtained by standard classifiers but their performance improvement could have reached a limit. Fusion, combination and ensemble of classifiers could represent a new step towards better emotion recognition systems.

This chapter aims to provide a comprehensive review on emotional speech recognition. The chapter is organized as follows. Section 2 describes the frameworks used for SER. Section 3 gives an overview of the types of databases. Section 4 presents the acoustic characteristics of emotions. Section 5 presents feature extraction and classification. Section 6 discusses the applications of emotion recognition. Section 7 presents concluding remarks.

2. Basic framework for emotional recognition

The input files are speech signals. Fig.1 gives the basic framework of emotional speech recognition. The feature extraction script extracts the features that represent global statistics. In the Post-processing step, the interface problem between the script for feature extraction and the feature selection technique can be solved. Then feature selection eliminates irrelevant features that hinder the recognition rates. It lowers the input dimensionality and saves the computational time. Distribution models like GMMs are trained using the most discriminative aspects of the feature. The classifiers distinguish the types of emotion.

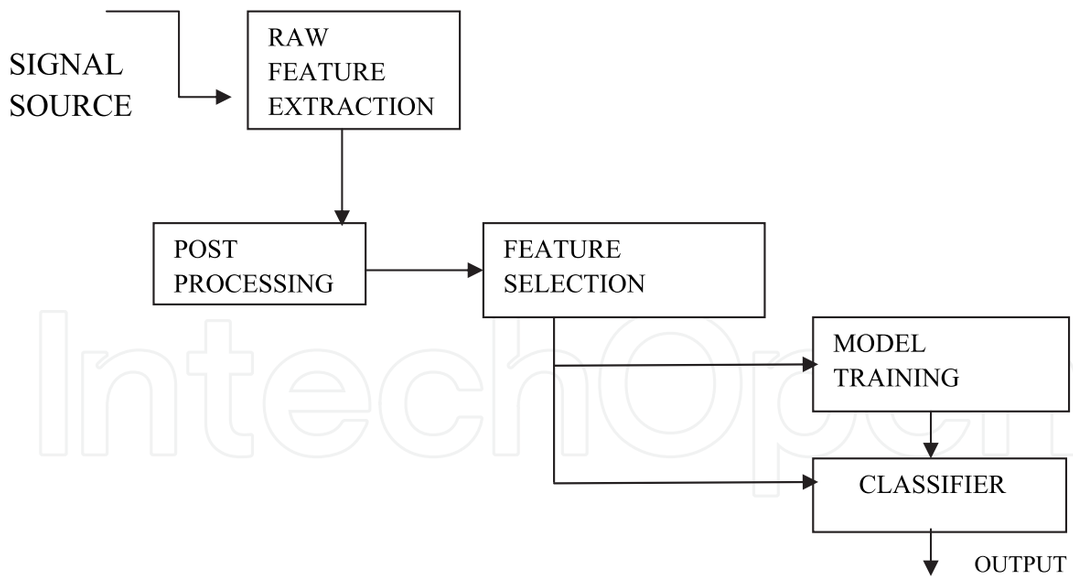


Fig. 1. Basic framework of SER

Bio signals such as ECG, EEG,GSR, face and body images are an interesting alternative to detect emotional states. Fig 2 discusses the mechanism of emotion recognition using these bio signals.

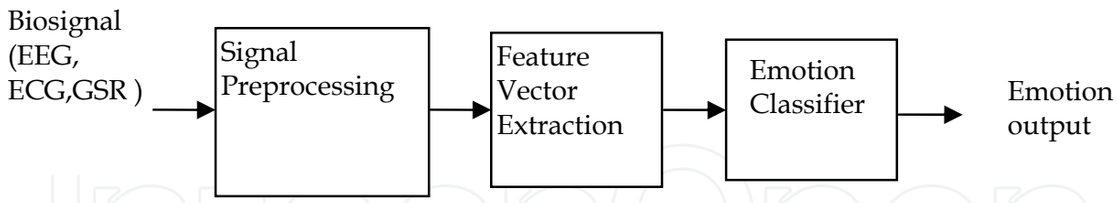


Fig. 2. Framework for emotion recognition using EEG,ECG,GSR signals

EEG is one of the most useful bio signals that detect true emotional state of human. The signal is recorded using the electrodes which measure the electrical activity of the brain. The recorded EEG data is first preprocessed to remove serious and obvious motion artifacts. Then the features are extracted from the raw signal using some feature extraction techniques like discrete wavelet transform, statistical based analysis etc. After the extraction the emotion classifier use the emotion classification techniques like Fuzzy C-Means, Quadratic Discriminant Analysis etc. to classify the different emotions of human.

ECG is recorded using ECG sensor .The signals are preprocessed using low pass filter at 100HZ. Then, features are extracted from the preprocessed signal by continuous wavelet transform (CWT) or discrete wavelets transform (DWT). Feature selection is done using Tabu Search Algorithm (TS), Simba algorithm etc. The selected feature is fed into classifier (fisher or K-Nearest Neighbor (KNN) classifier) to identify the type of emotion.

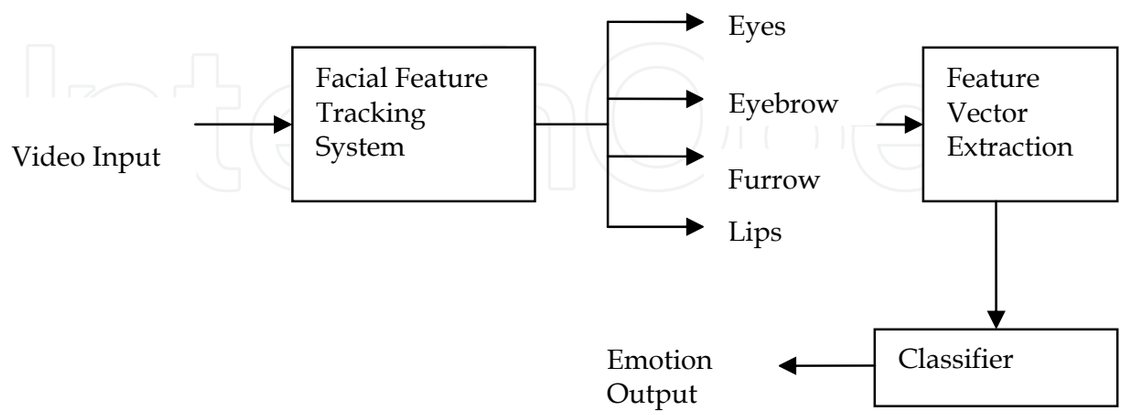


Fig. 3. Framework for Facial Emotional Recognition

Galvanic Skin Response is the measure of skin conductivity. There is a correlation between GSR and the arousal state of body. In the GSR emotional recognition system, the GSR signal is physiologically sensed and the feature is extracted using Immune Hybrid Particle Swarm Optimization (IH-PSO). The extracted features are classified using neural network classifier to identify the type of emotion.

In the facial emotion recognition the facial expression of a person is captured as a video and it is fed into the facial feature tracking system. Fig 3 gives a basic framework of facial emotional recognition.In facial feature tracking system, facial feature tracking algorithms such as Wavelets, Dual-view point-based model etc. are applied to track eyes, eyebrows, furrows and lips to collect all its possible movements. Then the extracted features are fed into classifier like Naïve Bayes , TAN or HMM to classify the type of emotion.

3. Emotional speech database

There should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. According to some studies the following are the most relevant factors to be considered:

- Real-world emotions or acted ones
- Who utters the emotions
- How to simulate the utterances
- Balanced utterances or unbalanced utterances
- Utterances are uniformly distributed over emotions

Most of the developed emotional speech databases are not available for public use. Thus, there are very few benchmark databases that can be shared among researchers. Most of the databases share the following emotions: anger, joy, sadness, surprise, boredom, disgust, and neutral.

Types of DB

At the beginning of the research on automatic speech emotion recognition, acted speech was used and now it shifts towards more realistic data. The databases that are used in SER are classified into 3 types. Fig 4 briefs the types of databases. Table 1 gives a detailed list of speech databases.

Type 1 is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion, e.g. DES, EMO-DB.

Type 2 is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come from real-life applications for example call-centers.

Type 3 is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated.

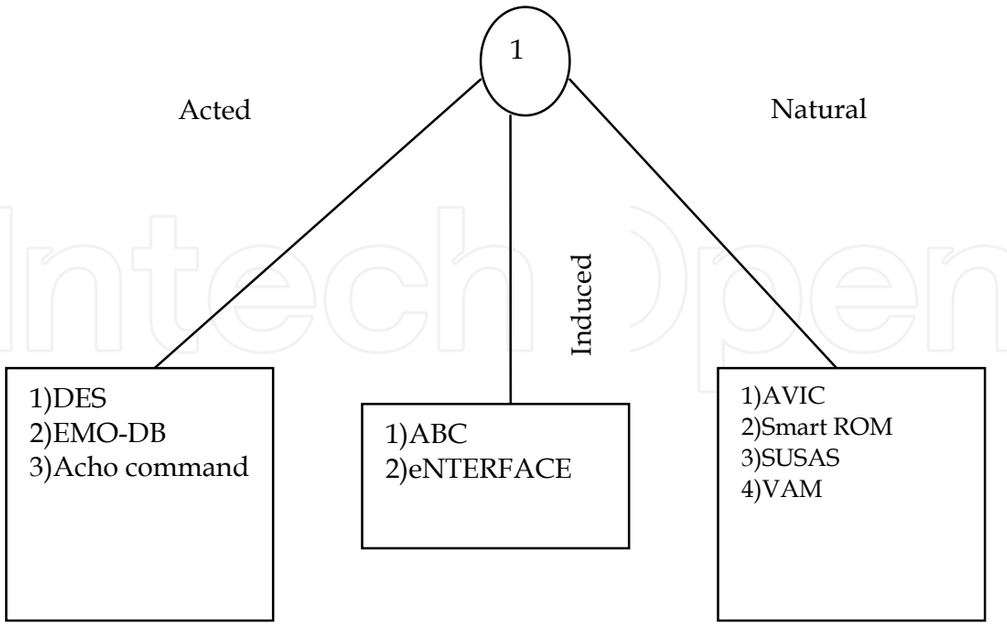


Fig. 4. Types of databases

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
1	ABC (Airplane Behaviour Corpus)	Total=8 Age=25–48 years Time=8.4 s/431 clips	Nature=Acted purpose= Transport surveillance Language=German Mode=Audio-Visual	Aggressive, Cheerful,Intoxicated, Nervous, Neutral, Tired	Publically Available=No Detection of Security Related Affect and Behaviour in Passenger Transport
2	EMO(Berlin Emotional Database)	Total=10 Male = 5 Female = 5	Nature=Acted Purpose =General Language =German Mode=Audio	Anger, Boredom, Disgust, Fear, Joy, Neutral, Sadness	Publically Available=Yes http://pascal.kgw.tu-berlin.de/emodb/docu/#download
3	SUSAS(Speech Under Stimulated and Actual Stress)	Total=32 Male = 19 Female= 13 Age=22-76 years	Nature= Induced Language =English Purpose=Aircraft Mode=Audio	Fear ,High Stress, Medium Stress, Neutral	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78
4	AVIC(Audiovisual Interest Corpus)	Total=21 Male =11 Female = 10	Nature= Natural Language =English Mode=Audio-Visual		http://webcache.googleusercontent.com/search?hl=en&start=10&q=cache:yp-tULzKJRwJ:http://citeseerx.ist.ps

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
					u.edu/viewdoc/download?doi=10.1.1.65.9121&rep=rep1&type=pdf+audiovisual+interest+speech+database&ct=clnk
5	SAL(Sensitive Artificial Listener)	Total=4 Female=2 Male=2 Time=20min/ speaker	Nature= Natural Purpose= Human-Computer conversation Language =English Mode=Audio-Visual		Publically Available=No http://emotion-research.net/toolbox/toolboxdatabase.2006-09-26.5667892524
6	Smartkom	Total =224 Time=4.5 min /person	Nature= Natural Purpose= Human-Computer conversation Language =German Mode=Audio-Visual	Neutral, Joy, Anger, Helplessness, Pondering, Surprise, Undefinable	Publically Available=No www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom Linguistic nature of material=Interactive discourse
7	VAM(Vera-Am-Mittag)	Total=47	Nature=Natural Language =German Mode=Audio-Visual	valence (negative vs. positive), activation (calm vs. excited) and dominance (weak vs. strong).	Publically Available=No http://emotion-research.net/download/vam
8	DES(Danish Emotional Database)	Total=4 Male = 2 Female = 2 Age=18 -58 years old	Nature=Acted Purpose =General Language =Danish Mode=Audio	Anger, Happiness, Neutral, Sadness, Surprise	Publically Available=Yes http://universal.elra.info/product_info.php?products_id=78
9	eNTERFACE	Total=42 Male = 34 Female = 8	Nature=Acted Purpose =General Language =English Mode=Audio-visual	Anger, Disgust, Fear, Joy, Sadness, Surprise	Publically Available=Yes Learning with synthesized speech for automatic emotion recognition
10	Groningen, 1996 ELRA corpus number S0020	Total=238	Nature=Acted Language = Dutch Mode=Audio		Publically Available=No www.elda.org/catalogue/en/speech/S0020.html Linguistic nature of material= subjects read too

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
					short text with many quoted sentences to elicit emotional speech
11	Pereira (Pereira, 2000a,b)	Total =2	Nature=Acted Language = English Mode=Audio	Anger(hot), Anger(cold), Happiness, Neutrality, Sadness	Emotional Speech Recognition:Resources,Feature and Method Linguistic nature of material= 2 utterances(1 emotionally neutral sentence,4 digit number) each repeated
12	Van Bezooijen (Van Bezooijen, 1984)	Total =8 Male = 4 Female = 4	Nature=Acted Language = Dutch Mode=Audio	Anger, Contempt, Disgust, Fear, Interest, Joy, Neutrality, Sadness, Shame, Surprise	Linguistic nature of material=4 semantically neutral phrases
13	Alter (Alter et al.,2000)	Total =1	Nature=Acted Language = German Mode=Audio	Anger(cold), Happiness, Neutrality	Emotional Speech Recognition:Resources,Feature and Method Linguistic nature of material=3 sentences,1 for each emotion(with appropriate content)
14	Abelin (Abelin and Allwood,2000)	Total =1	Nature=Acted Language = Swedish Mode=Audio	Anger, Disgust, Dominance, Fear, Joy, Sadness, Shyness, Surprise	A State of Art Review on Emotional Speech Database Linguistic nature of material=1 semantically neutral phrase
15	Polzin (Polzin and Waibel,2000)	Unspecified number of speakers	Nature=Acted Language = English Mode=Audio-Visual	Anger, Sadness, Neutrality(other emotions as well,but in insuffient numbers to be used)	Emotional Speech Recognition:Resources,Feature and Method Linguistic nature of material=sentence length segments taken

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
					from acted movies
16	Banse and scherer (Banse and scherer,1996)	Total =12 Male = 6 Female = 6	Nature=Induced Language =German Mode=Audio-Visual	Anger(hot), Anger(cold), Anxiety, Boredom, Contempt, Disgust, Elation, Fear(panic), Happiness, Interest, Pride, Sadness, Shame	Linguistic nature of material=2 semantically neutral sentences(non-sense sentences composed of phonemes from Indo-European languages)
17	Mozziconacci (Mozziconacci 1998)	Total =3	Nature=Induced Language =Dutch Mode=Audio	Anger, Boredom, Fear, Disgust, Guilt, Happiness, Haughtiness, Indignation, Joy, Neutrality, Rage, Sadness, Worry	Linguistic nature of material=8 semantically neutral sentences(each repeated 3 times)
18	Iriondo et al. (Iriondo et al., 2000)	Total =8	Nature=Induced Language =Spanish Mode=Audio	Desire, Disgust, Fury, Fear, Joy, Surprise, Sadness	Emotional Speech Recognition:Resources,Feature and Method ,Linguistic nature of material=paragraph length passages(20-40mms each)
19	McGilloway (McGilloway,1997;Cowie and Douglas-Cowie,1996)	Total =40	Nature=Induced Language =English Mode=Audio	Anger, Fear, Happiness, Neutrality, Sadness	Linguistic nature of material= paragraph length passages
20	Belfast structured database	Total =50	Nature=Induced Language =English Mode=Audio	Anger, Fear, Happiness, Neutrality, Sadness	Linguistic nature of material= paragraph length passages written in first person
21	Amir et al. (Amir et al.,2000)	Total=61(60 Hebrew speakers and 1 Russian speaker)	Nature = Induced Language=Hebrew, Russian Mode=Audio	Anger, Disgust, Fear, Joy, Neutrality, Sadness	Linguistic nature of material=non-interactive discourse
22	Femandez et al.(Femandez and Picard,2000)	Total=4	Nature=Induced Language =English Mode=Audio	Stress	Linguistic nature of material=numerical answers to mathematical questions
23	Tolkmitt and Scherer	Total =60 Male = 33 Female	Nature=Induced Language =German	Stress(both cognitive and emotional)	Emotional Speech

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
	(Tolkmitt and Scherer,1986)	=27	Mode=Audio		Recognition:Resources,Feature and Method Linguistic nature of material=subjects made 3 vocal responses to each slide within a forty seconds presentation period-a numerical answer followed by 2 short statements. The start of each was scripted and subjects filled in the blank at the end.
24	Reading-Leeds database (Greasley et al.,1995;Roach et al.,1998)	Time=264 min	Nature=Natural Language =English Mode=Audio		Automated Extraction Of Annotation Data From The Reading/Leeds Emotional Speech Corpus Speech Research Laboratory,University of Reading, Reading, RG1 6AA, UK Linguistic nature of material=unscripted interactive discourse
25	Belfast natural database (Douglas-Cowie et al., 2000)	Total =125 Male = 31 Female =94	Nature=Natural Language =English Mode=Audio-Visual	Wide range	Publically available=no http://www.idiap.ch/mmm/corpora/emotion-corpus Linguistic nature of material= unscripted interactive discourse

S.No	Corpus Name	No.of Subjects (Total, Male, female and age and time & days taken)	Nature(Acted/Natural/ Induced and purpose, Language& mode)	Types of Emotions(Anger, disgust, fear, joy, sad, etc)	Publicably Available(Yes/No) and URL
26	Geneva Airport Lost Luggage Study (Scherer and Ceschi,1997, 2000)	Total =109	Nature=Natural Language =Mixed Mode=Audio-Visual	Anger, Good humour, Indifference, Stress, Sadness	http://www.unige.ch/fapse/emotion/demo/TestAnalyst/GERG/apache/htdocs/index.php Linguistic nature of material= unscripted interactive discourse
27	Chung (Chung,2000)	Total =77 (61 Korean speakers,6 American speakers)	Nature=Natural Language =Korean, English Mode=Audio-Visual	Joy, Neutrality, Sadness(distress)	Linguistic nature of material= interactive discourse
28	France et al.(France et al.,2000)	Total =115 Male = 67 Female =48	Nature=Natural Language =English Mode=Audio	Depression, Neutrality, Suicidal state	Publically Available=no http://emotion-research.net/Members/admin/test/?searchterm=France%20et%20al.(France%20et%20al.,2000) Linguistic nature of material= interactive discourse
29	Slaney and McRoberts (1998) or Breazeal (2001)	Total =6	Nature=Acted Language =English, Japanese Purpose=pet robot Mode=Audio	Joy, Sadness, Anger, Neutrality	Publically Available=no
30	FAU Aibo Database	Total=26 children Male=13 Female=13	Nature=Natural Language =German Purpose=pet robot	Anger, Emphatic, Neutral, Positive, and Rest	Publically Available=no http://www5.cs.fau.de/de/mitarbeiter/steidl-stefan/fau-aibo-emotion-corpus/
31	SALAS database	Total=20	Nature=Induced Language =English Mode=Audio-Visual	Wide range	Publically Available=no http://www.image.ntua.gr/ermis/ IST-2000-29319, D09 Linguistic nature of material= interactive discourse

Table 1. List of emotional speech databases

4. Acoustic characteristics of emotions in speech

The prosodic features like pitch, intensity, speaking rate and voice quality are important to identify the different types of emotions. In particular pitch and intensity seem to be correlated to the amount of energy required to express a certain emotion. When one is in a state of anger, fear or joy; the resulting speech is correspondingly loud, fast and enunciated with strong high-frequency energy, a higher average pitch, and wider pitch range, whereas with sadness, producing speech that is slow, low-pitched, and with little high-frequency energy. In Table 2, a short overview of acoustic characteristics of various emotional states is provided.

EMOTIONS / CHARACTERISTICS	JOY	ANGER	SADNESS	FEAR	DISGUST
Pitch mean	High	very high	very low	very high	very low
Pitch range	High	high	Low	High	high-male low-female
Pitch variance	High	very high	Low	very high	Low
Pitch contour	incline	decline	Decline	Incline	Decline
Intensity mean	High	very high-male high-female	Low	medium/ high	Low
Intensity range	High	high	Low	High	Low
Speaking Rate	High	low-male high-female	high-male low-female	High	very low-male low-female
Transmission Durability	Low	low	High	Low	High
Voice Quality	modal/ tense	Sometimes breathy; Moderately blaring timbre	Resonant timbre	Falsetto	Resonant timbre

Table 2. Acoustic Characteristics of Emotions

5. Feature extraction and classification

The collected emotional data usually contain noise due to the background and "hiss" of the recording machine. The presence of noise will corrupt the signal, and make the feature extraction and classification less accurate. Thus preprocessing of speech signal is very much required. Preprocessing also reduces the variability.

Normalization is a preprocessing technique that eliminates speaker and recording variability while keeping the emotional discrimination. Generally 2 types of normalization techniques are performed they are energy normalization and pitch normalization. Energy normalization: the speech files are scaled such that the average RMS energy of the neutral

reference database and the neutral subset in the emotional databases are the same for each speaker. This normalization is separately applied for each subject in each database. The goal of this normalization is to compensate for different recording settings among the databases. Pitch normalization: the pitch contour is normalized for each subject (speaker-dependent normalization). The average pitch across speakers in the neutral reference database is estimated. Then, the average pitch value for the neutral set of the emotional databases is estimated for each speaker.

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still the data with sufficient accuracy.

Although significant advances have been made in speech recognition technology, it is still a difficult problem to design a speech recognition system for speaker-independent, continuous speech. One of the fundamental questions is whether all of the information necessary to distinguish words is preserved during the feature extraction stage. If vital information is lost during this stage, the performance of the following classification stage is inherently crippled and can never measure up to human capability. Typically, in speech recognition, we divide speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage. For example, for the case of dynamic time warping (DTW), this sequence of feature vectors is compared with the reference data set. For the case of hidden Markov models (HMM), vector quantization may be applied to the feature vectors which can be viewed as a further step of feature extraction. In either case, information loss during the transition from speech signals to a sequence of feature vectors must be kept to a minimum. There have been numerous efforts to develop good features for speech recognition in various circumstances.

The most common speech characteristics that are extracted are categorized in the following groups:

Frequency characteristics

- Accent shape – affected by the rate of change of the fundamental frequency.
- Average pitch – description of how high/low the speaker speaks relative to the normal speech.
- Contour slope – describes the tendency of the frequency change over time, it can be rising, falling or level.
- Final lowering – the amount by which the frequency falls at the end of an utterance.
- Pitch range – measures the spread between maximum and minimum frequency of an utterance.
- Formant-frequency components of human speech
- MFCC-representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.
- Spectral features- measures the slope of the spectrum considered.

Time-related features

- Speech rate – describes the rate of words or syllables uttered over a unit of time
- Stress frequency – measures the rate of occurrences of pitch accented utterances
- Energy- Instantaneous values of energy
- Voice quality- jitter and shimmer of the glottal pulses of the whole segment.

Voice quality parameters and energy descriptors

- Breathiness – measures the aspiration noise in speech
- Brilliance – describes the dominance of high Or low frequencies In the speech
- Loudness – measures the amplitude of the speech waveform, translates to the energy of an utterance
- Pause Discontinuity – describes the transitions between sound and silence
- Pitch Discontinuity – describes the transitions of fundamental frequency.

Durational pause related features : The duration features include the chunk length, measured in seconds, and the zero-crossing rate to roughly decode speaking rate. Pause is obtained as the proportion of non-speech to the speech signal calculated by a voice activity detection algorithm

Zipf features used for a better rhythm and prosody characterization.

Hybrid pitch features combines outputs of two different speech signal based pitch marking algorithms (PMA)

Feature selection determines which features are the most beneficial because most classifiers are negatively influenced by redundant, correlated or irrelevant features. Thus, in order to reduce the dimensionality of the input data, a feature selection algorithm is implemented to choose the most significant features of the training data for the given task. Alternatively, a feature reduction algorithm like principal components analysis (PCA) and Sequential Forward Floating Search (SFFS) can be used to encode the main information of the feature space more compactly.

Most research on SER has concentrated on feature-based and classification-based approaches. Feature-based approaches aim at analyzing speech signals and effectively estimating feature parameters representing human emotional states. The classification-based approaches focus on designing a classifier to determine distinctive boundaries between emotions. The process of emotional speech detection also requires the selection of a successful classifier which will allow for quick and accurate emotion identification. Currently, the most frequently used classifiers are linear discriminant classifiers (LDC), k-nearest neighbor (k-NN), Gaussian mixture model (GMM), support vector machines (SVM), decision tree algorithms and hidden Markov models (HMMs).Various studies showed that choosing the appropriate classifier can significantly enhance the overall performance of the system.

The list below gives a brief description of each algorithm:

LDC: A linear classifier uses the feature values to identify which class (or group) it belongs to by making a classification decision based on the value of a linear combination of the feature values .They are usually presented to the system in a vector called a feature vector.

k-NN: Classification happens by locating the instance in feature space and comparing it with the k nearest neighbors (training examples) and labeling the unknown feature with the same class label as that of the located (known) neighbor. The majority vote decides the outcome of class labeling.

GMM: A model of the probability distribution of the features measured in a biometric system such as vocal-tract related spectral features in a speaker recognition system. It is used for representing the existence of sub-populations, which is described using the mixture distribution, within the overall population.

SVM : It is a binary classifier to analyze the data and recognize the patterns for classification and regression analysis.

Decision tree algorithms: work based on following a decision tree in which leaves represent the classification outcome, and branches represent the conjunction of subsequent features that lead to the classification.

HMMs: It is a generalized model in which the hidden variables control the components to be selected. The hidden variables are related through the Markov process. In the case of emotion recognition, the outputs represent the sequence of speech feature vectors, which allow the deduction of states' sequences through which the model progressed. The states can consist of various intermediate steps in the expression of an emotion, and each of them has a probability distribution over the possible output vectors. The states' sequences allow us to predict the emotional state which we are trying to classify, and this is one of the most commonly used techniques within the area of speech affect detection.

Boostexter: an iterative algorithm that is based on the principle of combining many simple and moderately inaccurate rules into a single, highly accurate rule. It focuses on text categorization tasks. An advantage of Boostexter is that it can deal with both continuous-valued input (e.g., age) and textual input (e.g., a text string).

6. Applications

Emotion detection is a key phase in our ability to use users' speech and communications as a source of important information on users' needs, desires, preferences and intentions. By recognizing the emotional content of users' communications, marketers can customize offerings to users even more precisely than ever before .This is an exciting innovation that is destined to add an interesting dimension to the man-machine interface, with unlimited potential for marketing as well as consumer products, transportation, medical and therapeutic applications, traffic control and so on.

Intelligent Tutoring System: It aims to provide intervention strategies in response to a detected emotional state, with the goal being to keep the student in a positive affect realm to maximize learning potential. The research follows an ethnographic approach in the determination of affective states that naturally occur between students and computers. The multimodal inference component will be evaluated from audio recordings taken during classroom sessions. Further experiments will be conducted to evaluate the affect component and educational impact of the intelligent tutor.

Lie Detection: Lie Detector helps in deciding whether someone is lying or not. This mechanism is used particularly in areas such as Central Bureau of Investigation for finding out the criminals, cricket council to fight against corruption. **X13-VSA PRO Voice Lie Detector 3.0.1 PRO** is an innovative, advanced and sophisticated software system and a fully computerized voice stress analyzer that allows us to detect the truth instantly.

Banking: The ATM will employ speaker recognition and authentication if needed "to ensure higher security level while accessing to confidential data." In other words, the unique deployment of combining speech recognition, speaker recognition and emotion detection is not designed to be spooky or invasive. "It is just one more step forward the creation of humanlike systems that speak to the clients, understand and recognize a speaker". What's different is the incorporation of emotion detection in the enrollment process, which is probably a very good idea if enrollments are going to be conducted without human assistance or supervision. The machine will be able to talk with the prospective enrollee (and later on the client) and will be able to authenticate his or her unique voiceprint while, at the same time, test voice levels for signs of nervousness, anger, or deceit.

In-Car Board System: An in-car board system shall be provided with information about the emotional state of the driver to initiate safety strategies, initiatively provide aid or resolve errors in the communication according to the driver's emotion.

Prosody in Dialog System: We investigate the use of prosody for the detection of frustration and annoyance in natural human-computer dialog. In addition to prosodic features, we examine the contribution of language model information and speaking "style". Results show that a prosodic model can predict whether an utterance is neutral versus "annoyed or frustrated" with an accuracy on par with that of human interlobular agreement. Accuracy increases when discriminating only "frustrated" from other utterances, and when using only those utterances on which labelers originally agreed. Furthermore, prosodic model accuracy degrades only slightly when using recognized versus true words. Language model features, even if based on true words, are relatively poor predictors of frustration.

Emotion Recognition in Call Center: Call-centers often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can often lead to customer discontent, loss of business and in extreme cases, general customer unrest where a large amount of customers move to a competitor. It is therefore important for call-centers to take note of isolated disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied.

A system was designed to monitor recorded customer messages and provide an emotional assessment for more effective call-back prioritization. However, this system only provided post-call classification and was not designed for real time support or monitoring. Nowadays the systems are different because it aims to provide a real-time assessment to aid in the handling of the customer while he or she is speaking. Early warning signs of customer frustration can be detected from pitch contour irregularities, short-time energy changes, and changes in the rate of speech.

Sorting of Voice Mail: Voicemail is an electronic system for recording and storing of voice messages for later retrieval by the intended recipient. It would be a potential application to

sort the voice mail according to the emotion of the person's voice recorded. It will help to respond to the caller appropriately.

Computer Games: Computer games can be controlled through emotions of human speech. The computer recognizes human emotion from their speech and compute the level of game (easy, medium, hard). For example, if the human speech is in form of aggressive nature then the level becomes hard. Suppose if the human is too relaxed the level becomes easy. The rest of emotions come under medium level.

Diagnostic Tool By Speech Therapists: Person who diagnosis and treats variety of speech, voice, and language disorders is called a Speech Therapist. By understanding and empathizing emotional stress and strains the therapists can know what the patient is suffering from. The software used for recording and analyzing the entire speech is icSpeech. The use of speech communication in healthcare is to allow the patient to describe their health condition to the best of their knowledge. In clinical analysis, human emotions are analyzed based on features related to prosodics, the vocal tract, and parameters extracted directly from the glottal waveform. Emotional expressions can be referred by vocal affect extracted from the human speech.

Robots: Robots can interact with people and assist them in their daily routines, in common places such as homes, super markets, hospitals or offices. For accomplishing these tasks, robots should recognize the emotions of the humans to provide a friendly environment. Without recognizing the emotion, the robot cannot interact with the human in a natural way.

7. Conclusion

The process of speech emotion detection requires the creation of a reliable database, broad enough to fit every need for its application, as well as the selection of a successful classifier which will allow for quick and accurate emotion identification. Thirty-one emotional speech databases are reviewed. Each database consists of a corpus of human speech pronounced under different emotional conditions. A basic description of each database and its applications is provided. And the most common emotions searched for in decreasing frequency of appearance are anger, sadness, happiness, fear, disgust, joy, surprise, and boredom. The complexity of the emotion recognition process increases with the amount of emotions and features used within the classifier. It is therefore crucial to select only the most relevant features in order to assure the ability of the model to successfully identify emotions, as well as increasing the performance, which is particularly significant to real-time detection. SER has in the last decade shifted from a side issue to a major topic in human computer interaction and speech processing. SER has potentially wide applications. For example, human computer interfaces could be made to respond differently according to the emotional state of the user. This could be especially important in situations where speech is the primary mode of interaction with the machine.

8. References

[1] Zhihong Zeng, Maja Pantic I. Roisman, and Thomas S. Huang, 'A Survey of Affect Recognition Methods: Audio,Visual, and Spontaneous Expressions', IEEE

Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 1, pp.39-58, January 2009.

[2] Panagiotis C. Petrantonakis , and Leontios J. Hadjileontiadis, ' Emotion Recognition From EEG Using Higher Order Crossings', IEEE Trans. on Information Technology In Biomedicine, Vol. 14, No. 2,pp.186-197, March 2010.

[3] Christos A. Frantzidis, Charalampos Bratsas, et al 'On the Classification of Emotional Biosignals Evoked While Viewing Affective Pictures: An Integrated Data-Mining-Based Approach for Healthcare Applications', IEEE Trans. on Information Technology In Biomedicine, Vol. 14, No. 2, pp.309-318,March 2010

[4] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, , and Jyh-Horng Chen, 'EEG-Based Emotion Recognition in Music Listening', IEEE Trans. on Biomedical Engineering, Vol. 57, No. 7, pp.1798-1806 , July 2010.

[5] Meng-Ju Han, Jing-Huai Hsu and Kai-Tai Song, A New Information Fusion Method for Bimodal Robotic Emotion Recognition, Journal of Computers, Vol. 3, No. 7, pp.39-47, July 2008

[6] Claude C. Chibelushi, Farzin Deravi, John S. D. Mason, 'A Review of Speech-Based Bimodal Recognition', IEEE Transactions On Multimedia, vol. 4, No. 1 ,pp.23-37,March 2002.

[7] Bjorn Schuller , Bogdan Vlasenko, Florian Eyben , Gerhard Rigoll , Andreas Wendemuth, 'Acoustic Emotion Recognition:A Benchmark Comparison of Performances', IEEE workshop on Automatic Speech Recognition and Understanding , pp.552-557, Merano,Italy, December 13-20,2009.

[8] Ellen Douglas-Cowie , Nick Campbell , Roddy Cowie , Peter Roach, 'Emotional Speech: Towards a New Generation Of Databases' , Speech Communication Vol. 40, pp.33–60 ,2003.

[9] John H.L. Hansen, 'Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition', Speech Communication, Special Issue on Speech Under Stress,vol. 20(1-2), pp. 151-170, November 1996.

[10] Carlos Busso, , Sungbok Lee, , and Shrikanth Narayanan, , 'Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection', IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 4, pp.582-596, May 2009.

[11] Nathalie Camelin, Frederic Bechet, Géraldine Damnati, and Renato De Mori, ' Detection and Interpretation of Opinion Expressions in Spoken Surveys', IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, No. 2, pp.369-381, February 2010.

[12] Dimitrios Ververidis , Constantine Kotropoulos, 'Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition',Elsevier Signal Processing, vol.88,issue 12,pp.2956-2970,2008

[13] K B khanchandani and Moiz A Hussain, 'Emotion Recognition Using Multilayer Perceptron And Generalized Feed Forward Neural Network', IEEE Journal Of Scientific And Industrial Research Vol.68, pp.367-371,May 2009

[14] Tal Sobol-Shikler, and Peter Robinson, 'Classification of Complex Information: Inference of Co-Occurring Affective States from Their Expressions in Speech', IEEE

Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 7, pp.1284-1297, July 2010

[15] Daniel Erro, Eva Navas, Inma Hernáez, and Ibon Saratxaga, 'Emotion Conversion Based on Prosodic Unit Selection' , IEEE Transactions On Audio, Speech And Language Processing, Vol. 18, No. 5, pp.974-983, July 2010

[16] Khiet P. Truong and Stephan Raaijmakers, 'Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features', MLMI 2008, LNCS 5237, pp. 161–172, 2008.

[17] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, 'Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture', IEEE International Conference on Acoustics, Speech, and Signal Processing, Quebec,Canada,17-21 May,2004

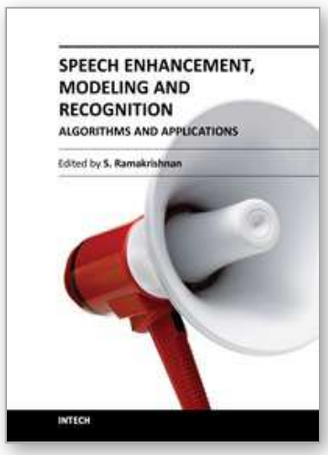
[18] Bjorn Schuller, Bogdan Vlasenko, Dejan Arsic, Gerhard Rigoll, Andreas Wendemuth, 'Combining Speech Recognition and Acoustic Word Emotion Models for Robust text-Independent Emotion Recognition', IEEE International Conference on Multimedia & Expo,Hannover,Germany,June 23-26,2008

[19] Wernhuar Tarng, Yuan-Yuan Chen, Chien-Lung Li, Kun-Rong Hsie and Mingteh Chen, 'Applications of Support Vector Machines on Smart Phone Systems for Emotional Speech Recognition', World Academy of Science, Engineering and Technology Vol.72, pp.106-113, 2010

[20] Silke Paulmann , Marc D. Pell , Sonja A. Kotz, 'How aging affects the recognition of emotional speech', Brain and Language Vol. 104, pp.262–269,2008

[21] Elliot Moore II, Mark A. Clements, , John W. Peifer, , and Lydia Weisser , 'Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech', IEEE Transactions On Biomedical Engineering,Vol. 55, No. 1, pp.96-107, January 2008.

[22] Yongjin Wang and Ling Guan, Recognizing Human Emotional State From Audiovisual Signals, IEEE Transactions on Multimedia, Vol. 10, No. 4, pp. 659-668, June 2008.



Speech Enhancement, Modeling and Recognition- Algorithms and Applications

Edited by Dr. S Ramakrishnan

ISBN 978-953-51-0291-5
Hard cover, 138 pages
Publisher InTech
Published online 14, March, 2012
Published in print edition March, 2012

This book on Speech Processing consists of seven chapters written by eminent researchers from Italy, Canada, India, Tunisia, Finland and The Netherlands. The chapters covers important fields in speech processing such as speech enhancement, noise cancellation, multi resolution spectral analysis, voice conversion, speech recognition and emotion recognition from speech. The chapters contain both survey and original research materials in addition to applications. This book will be useful to graduate students, researchers and practicing engineers working in speech processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

S. Ramakrishnan (2012). Recognition of Emotion from Speech: A Review, Speech Enhancement, Modeling and Recognition- Algorithms and Applications, Dr. S Ramakrishnan (Ed.), ISBN: 978-953-51-0291-5, InTech, Available from: http://www.intechopen.com/books/speech-enhancement-modeling-and-recognition-algorithms-and-applications/recognition-of-emotion-from-speech-a-review-

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen